Robust Speech Recognition Based on Dereverberation Parameter Optimization Using Acoustic Model Likelihood

AUTHOR(S):

Gomez, Randy; Kawahara, Tatsuya

Robust Speech Recognition Based on Dereverberation Parameter Optimization Using Acoustic Model Likelihood

Randy Gomez and Tatsuya Kawahara, *Senior Member, IEEE*

***Abstract*—Automatic speech recognition (ASR) in reverberant environments is a challenging task. Most dereverberation techniques address this problem through signal processing and enhances the reverberant waveform independent from the speech recognizer. In this paper, we propose a novel scheme to perform dereverberation in relation with the likelihood of the back-end ASR system. Our proposed approach effectively selects the dereverberation parameters, in the form of multiband scale factors, so that they improve the likelihood of the acoustic model. Then, the acoustic model is retrained using the optimal parameters. During the recognition phase, we implement additional optimization of the parameters. By using Gaussian mixture model (GMM), the process for selecting the scale factors become efficient. Moreover, we remove the dependency of the adopted dereverberation technique on the room impulse response (RIR) measurement, by using an artificial RIR generator and selecting based on the acoustic likelihood. Experimental results show significant improvement in recognition performance with the proposed method over the conventional approach.**

***Index Terms*—Automatic speech recognition (ASR), dereverberation, robustness.**

I. Introduction

REVERBERATION is a phenomenon that biases the speech signal due to reflections. The effect of reverberation drastically degrades the performance of the automatic speech recognition (ASR). Dereverberation technique is often employed as a speech enhancement tool to minimize the smearing effects prior to acoustic model training or input to ASR. Dereverberation techniques vary in different approaches. A technique based on inverse filtering [1], [2] performs deconvolution of the reverberant signal with an inverse filter. This can be done with the assumption that the room impulse response (RIR) is available. However, RIR measurement for every room is not practical for ASR applications. RIR also varies accordingly with the change in room conditions, such as speaker movement and location. A more sophisticated extension of inverse filtering is the use of the subspace method [3], [4] which estimates the inverse filter for blind deconvolution independent of the source characteristics.

Another interesting approach is based on probabilistic model of the speech source. The time-varying characteristics of the speech signal is incorporated in the derevereberation formulation, and optimization criterion is defined based on the probabilistic model of time-varying short-term speech characteristics. An example of the approach is a variational speech enhancement [5] which transforms the dereverberation problem into Bayes-optimal signal estimation. The speech signal is reconstructed by estimating the speech model parameters and filter coefficients through an expectation–maximization (EM) algorithm using a large database of clean speech. Another interesting method involving the probabilistic model formulation is proposed in [6], where dereverberation is formulated as a maximum-likelihood problem using a hill-climbing method and the speech signal is recovered by transforming the observed reverberant signal into one that is probabilistically more like non-reverberant speech. The method can be fine-tuned using hidden Markov model (HMM) [7].

We have also previously presented a dereverberation approach based on multiband spectral subtraction [8]–[10] which was implemented so that the scaling factors are optimized based on minimum mean square error (MMSE) criterion to effectively remove the late reflection components. A similar approach is proposed in [11], which uses a multistep linear prediction in estimating the late reflection.

These approaches work well in enhancing the reverberant signal in the waveform level. While the perceptual quality is improved together with the signal-to-noise ratio, the optimization used in these dereverberation approaches have no direct relation to the ASR system. The main objective of this paper is to develop an effective way of optimizing dereverberation so that it is inclined to improving the performance of the speech recognizer, while using spectral subtraction as the dereverberation framework. This kind of approach, where front-end speech processing is optimized for speech recognition is shown to be effective with promising results in microphone array applications [12], [13] and in vocal tract length normalization (VTLN) [14]–[16].

The proposed approach is embodied with four steps. First, we introduce a supervised optimization of the dereverberation parameters based on the likelihood of the speech recognizer instead of using MMSE [8]–[10]. Since this is done offline with the training data readily available, supervised optimization will result in a better performance. Second, we apply the optimized

Manuscript received November 13, 2009; revised February 26, 2010; accepted May 21, 2010. Date of current version August 13, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomohiro Nakatani.

The authors are with the ACCMS, Kyoto University, Kyoto 606-8501, Japan (e-mail: randy-g@ar.media.kyoto-u.ac.jp; kawahara@ar.media.kyoto-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2052610

dereverberation parameters in the acoustic model training phase. Third, we implement an unsupervised optimization of the dereverberation parameters during the speech recognition phase. Gaussian mixture model (GMM) is employed for fast decoding of utterances without transcription for each choice of the parameters. This will further minimize the mismatch between training and testing conditions. Moreover, an RIR generator is employed in order to estimate the reverberation time T_{60} of the room based on the likelihood criterion. This removes the dependence on the RIR measurement which is required in many conventional methods [1], [2], [8]–[10].

We evaluate the effectiveness of the proposed approach using real reverberant data together with artificially-generated reverberant data. The organization of the paper is as follows. In Section II, we show an overview of the multiband spectral subtraction, which is adopted in this paper as the dereverberation scheme. In Section III, we present the optimization of the dereverberation parameters based on the acoustic likelihood, followed by the optimization during decoding in Section IV. In Section V, we discuss the method of approximating RIR, followed by the experimental evaluation in Section VI. We will conclude this paper in Section VII.

II. Dereverberation Based on Spectral Subtraction

In this section, we introduce the principle of the dereverberation approach we intend to optimize. First, the spectral subtraction commonly used in denoising is introduced. Then, we discuss its extension to the multiband dereverberation based on MMSE.

A. Spectral Subtraction

Spectral subtraction (SS) was originally used as a noise suppression technique introduced in seminal works [17]. Several methods have been proposed and reported to be effective in many applications [18]–[20]. Here, we will present the simplest form of spectral subtraction based on a single band. The noisy signal can be modeled as

$$x(n) = s(n) + e(n) \tag{1}$$

where $s(n)$ is the clean speech signal corrupted by an additive noise $e(n)$ at sample index n. In a real scenario, we have access to this noisy observation $x(n)$. From this, we can estimate $\hat{e}(n)$ for the additive noise $e(n)$ and $\hat{s}(n)$ for $s(n)$.

The power spectra of $x(n)$ and $\hat{e}(n)$ are estimated on a frame-by-frame basis using short-term Fourier transform. The frequency domain representation of $x(n)$ is

$$X(f,t) = |X(f,t)|\, e^{\mathbf{j}\phi_X(f,t)}$$

and its power spectrum is $|X(f,t)|^2$, where f denotes frequency, t as period, and $\phi_X(f,t)$ as the phase. Considering (1), and assuming that the signal and noise are uncorrelated, the power spectrum of the noisy signal can be written as

$$|X(f,t)|^2 \approx |S(f,t)|^2 + |E(f,t)|^2 \tag{2}$$

where $|S(f,t)|^2$ and $|E(f,t)|^2$ are, respectively, the power spectra of $s(n)$ and $e(n)$. Then, the estimate of the power spectrum of $\hat{s}(n)$ through spectral subtraction is

$$|\widehat{S(f,t)}|^2 = |X(f,t)|^2 - |\hat{E}(f,t)|^2. \tag{3}$$

Because of the estimation error, for some values of f and t, the noise power spectrum estimate may be larger than the power spectrum of the true noise, resulting in a negative estimate $|\widehat{S(f,t)}|^2$. For this reason, $\hat{s}(n)$ is obtained by taking the inverse short-term Fourier transform from the following $\hat{S}(f,t)$

$$\hat{S}(f,t) = \begin{cases} \sqrt{|\widehat{S(f,t)}|^2}\, e^{\mathbf{j}\phi_X(f,t)}, & \text{if } |\widehat{S(f,t)}|^2 > 0 \\ \beta |X(f,t)|\, e^{\mathbf{j}\phi_X(f,t)}, & \text{otherwise} \end{cases}$$

where β, the flooring parameter is used to correct the negative values of $|\widehat{S(f,t)}|^2$. Note that the phase $\phi_X(f,t)$ is derived from the noisy signal.

B. Spectral Subtraction for Dereverberation

The spectra of the reverberant signal is given as

$$X(f) \approx S(f)H(f) \tag{4}$$

where $X(f)$, $S(f)$, and $H(f)$ are the frequency components of the reverberant signal, clean speech signal and the RIR, respectively. The reverberation effect can be decomposed into early and late reflections. The early reflection is due to the direct signal and some reflections that occur at earlier time and can be treated as short-period noise. The late reflection, whose effect spans over frames can be treated as long-period noise. The RIR h can be expressed with early h_E and late h_L components as follows:

$$h_E(\tau) = \begin{cases} h(\tau), & \tau < \Gamma \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$h_L(\tau) = \begin{cases} h(\tau + \Gamma), & \tau \geq \Gamma \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where τ denotes the time sample. Equations (5) and (6) characterize both the short and long-period effects of the reverberant signal. We denote the STFT of the reverberant signal, clean speech signal, and RIR as $X(f,t)$, $S(f,t)$, and $H(f,t)$, respectively. Based on (4), the reverberant speech model expressed in terms of early and late reflections is approximated as

$$\begin{aligned} X(f,t) &\approx S(f,t)H(f,0) + \sum_{d=1}^{D} S(f,t-d)H(f,d) \\ &\approx X_E(f,t) + X_L(f,t) \end{aligned} \tag{7}$$

where $H(f,0)$ is the RIR effect to the speech signal $S(f,t)$ due to $h_E(t)$. We denote this as early reflection $X_E(f,t)$. The second term $X_L(f,t)$ referred to as the late reflection can be viewed as smearing of the clean speech by $H(f,d)$ which corresponds to the d frame-shift effect of the RIR due to $h_L(t)$. D is the number of frames over which the reverberation (smearing) has an effect. The early reflection is mostly addressed through cepstral mean normalization (CMN) in the ASR system as it

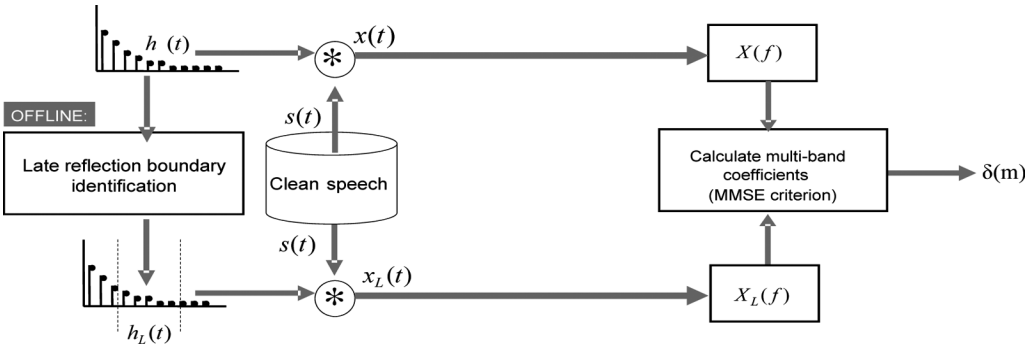

Fig. 1. Obtaining the values of the multiband coefficients offline using the clean utterances.

falls within the frame. In the spectral subtraction-based dereverberation, we are only interested in suppressing the effect of the late reflection. Theoretically, we can formulate to remove the entire reverberation effect, but robustness to the microphone-speaker location cannot be achieved since the early components $h_E(t)$ have high energy and is dependent on the distance between the microphone and the speaker, as explained in [8], [9] . Thus, for recovering the early reflection through spectral subtraction, we have (8), shown at the bottom of the page. Here, $|X(f,t)|^2$ and $|X_L(f,t)|^2$ are the power spectra of the reverberant signal and its late reflection respectively. Since the correlation of the speech signal decays over time, we can make the uncorrelated assumption for X_E and X_L as in (2) [21]. After performing spectral subtraction given in (8), we employ CMN to address the effects of early reflection.

C. Late Reflection Estimation

We note that estimating the late reflection requires the clean speech information as shown in the second term of (7). However, we do not have access to the clean speech in the actual scenario. Thus, we devise a scheme [8], [9], in a form of offline training, to approximate the late reflection using the reverberant speech as shown in Fig. 1. As we have access to the clean speech s in the training database, x and x_L are then generated by convolving s with the RIR h and its late components h_L, respectively. Next, the corresponding power spectral components $X(f)$ and $X_L(f)$ are calculated. For a given set of bands $\boldsymbol{B} = \{B_1, \ldots, B_M\}$, the coefficients $\boldsymbol{\delta}_{\text{MMSE}} = [\delta(1)_{\text{MMSE}}, \delta(2)_{\text{MMSE}}, \delta(m)_{\text{MMSE}}, \ldots, \delta(M)_{\text{MMSE}}]$ are determined by minimizing the mean squared error (MMSE) in each band m

$$E_m = \frac{1}{T}\sum_t \sum_{f \in B_m} |X_L(f,t) - \delta(m)(f,t)X(f,t)|^2 \qquad (9)$$

where ξ is the expectation operator. The minimization of error E_m is more effective by introducing multiple bands which give a finer treatment. Once the optimal $\delta(m)$ is found, we can substitute $X_L(f)$ with $\delta(m)X(f)$ and rewrite (8) into multiband spectral subtraction, as shown in (10) at the bottom of the page, where β is the flooring coefficient. In this work, we set $\beta = 0.001$. This process is referred to as MMSE-based spectral subtraction. The boundary of h_L, which is defined by Γ and D, was identified experimentally in our previous work [8], [9] .

III. Optimization of Dereverberation Parameters for Acoustic Model Training

The conventional approach adopts the MMSE criterion in deriving the scale parameters $\delta(m)$ for bands $m = 1, \ldots M$, which is subsequently used to process the reverberant signal prior to acoustic model training. We present two methods that optimize the dereverberation parameters jointly with acoustic model training. This principle is also applied during actual speech recognition which will be discussed in Section IV. Embedding the optimization process of the multiband scale factors in acoustic model training is not straightforward. Consider a single-band optimization given as

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{r=1}^{R} \max_{\delta} P(\mathbf{x}_r^{\delta}|\mathbf{w};\lambda) \qquad (11)$$

$$|X_E(f,t)|^2 = \begin{cases} |X(f,t)|^2 - |X_L(f,t)|^2, & \text{if } |X(f,t)|^2 - |X_L(f,t)|^2 > 0 \\ \beta|X(f,t)|^2, & \text{otherwise} \end{cases} \qquad (8)$$

$$|X_E(f,t)|^2 = \begin{cases} |X(f,t)|^2 - \delta(m)|X(f,t)|^2, & \text{if } |X(f,t)|^2 - \delta(m)|X(f,t)|^2 > 0 \\ \beta|X(f,t)|^2, & \text{otherwise} \end{cases} \qquad (10)$$

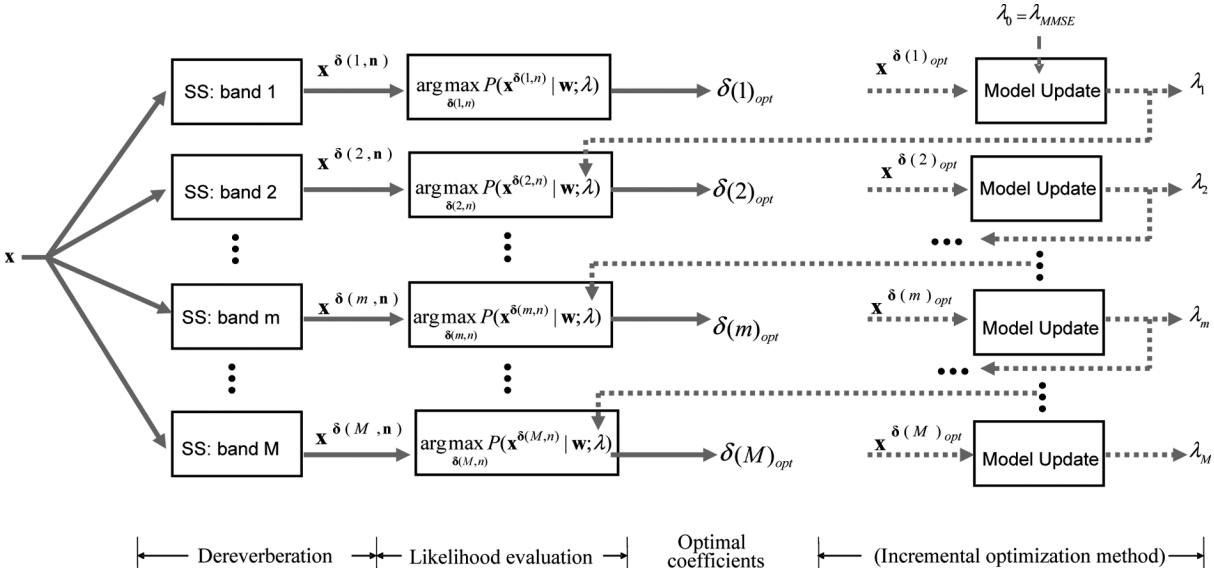

Fig. 2. Block diagram of the proposed optimization and acoustic model training.

where λ is unknown model parameters, δ is the single-band scale factor, and $\mathbf{w}$ is the word sequence. $\mathbf{x}_r^{\delta}$ is the rth training utterance processed by using the δ scale factor. The optimization of δ has to be carried out separately for each of the training utterance over the unknown model parameter λ, which is a difficult problem. The optimization problem is further complicated when expanded from a single band to multiple bands where each of the multiband scaling factors is to be optimized for each training utterances.

In this section, we present an alternative implementation of the optimization criterion in (11) and expand it to the multiband optimization. Prior to the proposed optimization, the reverberant data is processed with MMSE-based spectral subtraction (9) and a baseline acoustic model is trained with all the training utterances. We denote the observation data as $\mathbf{x}_r^{\boldsymbol{\delta}_{\mathrm{MMSE}}}$ which means that it is processed by the conventional MMSE-based multiband spectral subtraction, then a model can be estimated using them:

$$\lambda_{\mathrm{MMSE}} = \arg\max_{\lambda} \prod_{r=1}^{R} P(\mathbf{x}_r^{\boldsymbol{\delta}_{\mathrm{MMSE}}} | \mathbf{w}\lambda). \tag{12}$$

λ_{MMSE} is used in the optimization process discussed in the following subsections.

A. Batch Optimization Method

The proposed batch optimization of the multiband spectral subtraction is shown in Fig. 2. Initially, all the bands are set through the MMSE criterion $\boldsymbol{\delta}_{\mathrm{MMSE}} = [\delta(1)_{\mathrm{MMSE}}, \ \delta(2)_{\mathrm{MMSE}}, \delta(m)_{\mathrm{MMSE}}, \ldots, \delta(M)_{\mathrm{MMSE}}]$. Then, the particular band scale factor to be optimized is allowed to vary within a close neighborhood $n\triangle$, where $n = \pm 1 \cdots N$ and $\triangle = 0.02$ as the step size. The general expression of generating the sets of coefficients for the mth optimal scale factor in the batch optimization is given as (13), shown at the bottom of the page.

In (13), we opt to optimize each band independently, so only the band of interest is allowed to vary while the rest of the bands are kept to their respective MMSE values. Thus, we generate $2n$ different sets of coefficients denoted by $\boldsymbol{\delta}(m, n)$ for every band m. Spectral subtraction is executed using the sets of generated coefficients. The resulting data $\mathbf{x}^{\boldsymbol{\delta}}(m, n)$ are evaluated using the baseline acoustic model which is trained with data processed with the MMSE-based scaling factors, denoted as $\lambda = \lambda_{\mathrm{MMSE}}$. Then, the optimal scaling factor that gives the maximum likelihood is selected:

$$\delta(m)_{\mathrm{opt}} = \arg\max_{\boldsymbol{\delta}(m,n)} P(\mathbf{x}_r^{\boldsymbol{\delta}(m,n)} | \mathbf{w}; \lambda). \tag{14}$$

The whole process from spectral subtraction to likelihood evaluation is applied to all M bands independently $(m = 1, \ldots M)$. After all of the bands are optimized, the set of optimal spectral subtraction coefficients

$$\boldsymbol{\delta}_{\mathrm{opt}} = [\delta(1)_{\mathrm{opt}}, \ldots \delta(m)_{\mathrm{opt}}, \ldots, \delta(M)_{\mathrm{opt}}]$$

is used to process the reverberant data and retrain the acoustic model:

$$\lambda_{\mathrm{opt}} = \arg\max_{\lambda} \prod_{r=1}^{R} P(\mathbf{x}_r^{\boldsymbol{\delta}_{\mathrm{opt}}} | \mathbf{w}; \lambda).$$

Note that $\boldsymbol{\delta}_{\mathrm{opt}}$ is searched for every training utterance $\mathbf{x}_r$.

$$\boldsymbol{\delta}(m, n) = [\delta(1)_{\mathrm{MMSE}}, \ \delta(2)_{\mathrm{MMSE}}, \ldots, \delta(m)_{\mathrm{MMSE}} + n\triangle, \ldots, \delta(M)_{\mathrm{MMSE}}] \tag{13}$$

B. Incremental Optimization Method

We extend the batch optimization method. The additional process introduced is shown in dashed lines in Fig. 2. Right after the optimal coefficient of band 1 is found, the acoustic model is re-estimated using the updated spectral subtraction parameters. The general expression of generating the sets of candidate scale factors when optimizing the mth band scale parameter is also updated as shown in (15) at the bottom of the page, and the optimal scaling factor at the particular band that results in maximum likelihood is selected

$$\delta(m)_{\text{opt}} = \arg\max_{\boldsymbol{\delta}(m,n)} P(\mathbf{x}_r^{\boldsymbol{\delta}(m,n)}|\mathbf{w};\lambda_{m-1}).$$

Then, the utterances are processed with the updated scale factors expressed as (16) shown at the bottom of the page. The newly re-estimated model

$$\lambda_m = \arg\max_{\lambda_{m-1}} \prod_{r=1}^{R} P(\mathbf{x}_r^{\boldsymbol{\delta}(m)_{\text{opt}}}|\mathbf{w};\lambda_{m-1}) \tag{17}$$

is then used in the likelihood evaluation block for band $(m+1)$, and this process is iterated until the Mth band. λ_m is the mth re-estimated model trained using the utterances $\mathbf{x}_r^{\boldsymbol{\delta}(m)_{\text{opt}}}$ which is processed with the updated scale factors given in (16). The final model training is done as below

$$\lambda_{\text{opt}} = \arg\max_{\lambda_M} \prod_{r=1}^{R} P(\mathbf{x}_r^{\boldsymbol{\delta}_{\text{opt}}}|\mathbf{w};\lambda_M),$$

where

$$\boldsymbol{\delta}_{\text{opt}} = [\delta(1)_{\text{opt}}, \ldots \delta(m)_{\text{opt}}, \ldots, \delta(M)_{\text{opt}}]$$

and λ_M is the Mth retrained model in (17).

This approach, referred to as incremental optimization method, has the same principle with the batch method, except for the incremental updates of the HMM parameter λ in every band. In the batch method, we fixed $\lambda = \lambda_{\text{MMSE}}$ throughout all bands. The incremental re-estimation allows us to treat each band interdependently in a sequential manner as opposed to the batch optimization method where each band is treated independently.

IV. Fast Multiband Scale Factor Selection During Decoding

To compensate for the mismatch in the reverberant condition between training and testing, we implement additional optimization during speech recognition. After the training

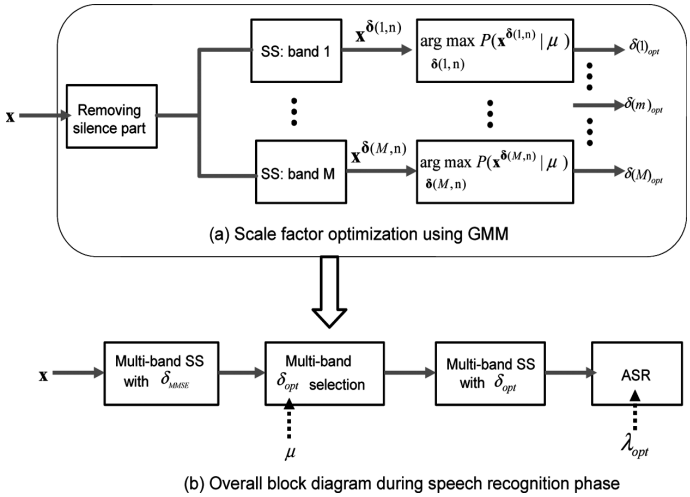

Fig. 3. Fast scale factor selection for speech recognition phase.

in Section III, a GMM denoted as μ with 64 components is trained using the dereverberated data processed with the optimal scaling factors $\boldsymbol{\delta}_{\text{opt}}$. This is a text-independent model which only captures the statistical information pertaining to the optimized multiband dereverberation parameters. A similar approach is reported in rapid speaker selection [25], [26] and also in VTLN training [16].

It is important that the optimization procedure in the decoding phase should be fast in order to execute real-time recognition. Fig. 3(a) shows the block diagram of the optimization technique, where the actual reverberant test data are processed with the multiband spectral subtraction. After the silence parts (low-energy segments) are removed, the utterance is fully evaluated with GMM for each choice of the scale factor. Subsequently, the scale factor that leads to the best likelihood is selected. We note that since a very simple model is used as opposed to HMM, the decoding is fast and efficient. In Fig. 3(b), we show the overall block diagram of the whole process from the scale parameter selection to the final ASR. In this figure, it is apparent that GMM is only used in the optimization process while HMM is used in the ASR.

V. Automatic Estimation of Reverberation Time (T_{60})

A number of dereverberation approaches rely on a readily available RIR measurement [1], [2] . An effective technique in measuring RIR is described in [22]. However, it is impractical to measure RIR for every room where the ASR system is deployed. In this section, we address the technique of modeling

$$\boldsymbol{\delta}(m,n) = [\delta(1)_{\text{opt}},\ \delta(2)_{\text{opt}}, \ldots,\ \delta(m)_{\text{MMSE}} + n\triangle, \ldots, \delta(M)_{\text{MMSE}}] \tag{15}$$

$$\boldsymbol{\delta}(m)_{\text{opt}} = [\delta(1)_{\text{opt}},\ \delta(2)_{\text{opt}}, \ldots,\ \delta(m)_{\text{opt}}, \ldots, \delta(M)_{\text{MMSE}}] \tag{16}$$

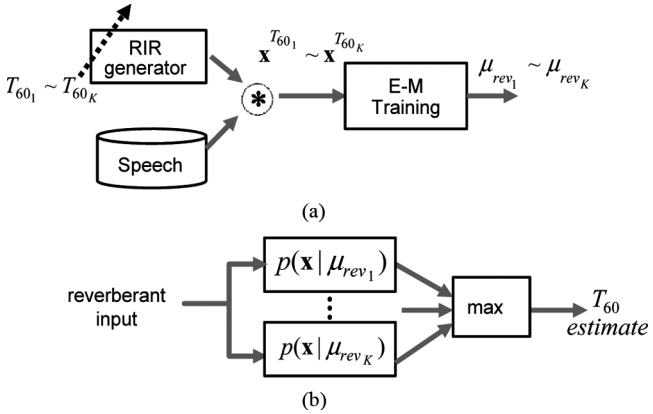

Fig. 4. Room impulse response approximation. (a) GMM setup with different T_{60} values. (b) T_{60} estimation based on acoustic likelihood.

the transmission in a reverberant room using an exponential decaying shape as introduced in [23]. Furthermore, we discuss the extension of this method to better fit our application.

Each phone HMM represents a short speech segment with a duration of 30–100 ms. Each state captures information about a distribution of spectral parameters. With this perspective, the HMMs' description of speech is of low resolution, compared to the RIR, with respect to time and frequency. Thus, for ASR applications, it may be sufficient to use an approximate RIR instead of the accurate RIR [23]. Existing studies suggest that ideally, the multiple reflections of sound can be described by a decaying acoustical energy, and the decay is best modeled by an exponential function [24]. Thus, the energy of the RIR is given as

$$h^2(l) \approx e^{(6\ \ln\ (10)/T_{60})\ l} \tag{18}$$

where l is the discrete time sample, and T_{60} is the reverberation time with a value ranging between 0.2–0.4 s. for smaller rooms and 0.4–0.8 s. for larger rooms [23].

To effectively identify T_{60}, we design a GMM-based T_{60} classifier as shown in Fig. 4(a). Reverberant speech data are synthetically generated $x^{T_{60_k}}$ with variable T_{60_k} values to train GMMs μ_{rev_k}. We use a large-mixture GMM (i.e., 256 mix.) to better capture the RIR characteristics. Fig. 4(b) shows the actual identification of T_{60} where the likelihood is calculated using the reverberant speech given all of the GMMs with different T_{60_k}. The corresponding T_{60} that results in the highest likelihood score is selected and from this, the RIR is estimated using (18).

In this paper, we experimentally set the step size of T_{60} to 30 ms covering from 100 ms to 1 s. Another approach in T_{60} estimation is presented in [23] where initial T_{60} estimate is allowed to vary at a certain step size and a maximum-likelihood search is conducted with the recognized HMM sequence. We compared the proposed method with [23] and found both achieve comparable performance in T_{60} estimation. The difference between our method and [23] is that, we use a simple GMM classifier to capture the characteristics of T_{60} while the latter is based on the realignment of HMMs.

The order of the system processes from training to testing is summarized as follows: prior to training, reverberation time

TABLE I
SYSTEM SPECIFICATION FOR TESTING

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCCs, 12-order ΔMFCCs, and 1-order ΔE
HMM	8256 Gaussian pdfs
Training data	JNAS convolved with measured RIR (T1) JNAS convolved with estimated RIR (T2)
Testing data	JNAS test set recorded in a real reverberant room ($\boldsymbol{E0}$) JNAS test set convolved with measured ($\boldsymbol{E1}$) and estimated ($\boldsymbol{E2}$) RIRs

T_{60} is estimated as discussed above, to generate the reverberant training data. Then, optimized dereverberation is implemented as described in Section III, which is followed by acoustic model training. During testing, the multiband scale factor is used for the test data as introduced in Section IV. When the system is used in the same room, where room acoustic does not vary much, robustness to mismatch between training and testing is achieved through the multiband scale factor selection (see Section IV) without the need of model retraining. In case where room acoustics vary severely (i.e., different T_{60}), we perform a model retraining (see Section III) to generate a matched acoustic model.

VI. EXPERIMENTAL EVALUATION

A. Training and Testing Data

For evaluation of the proposed method, we used the Japanese Newspaper Article Sentence (JNAS) corpus. The training database is composed of 100 male and female speakers with a total of approximately 60 hours of speech. The test set is composed 25 male and 25 female speakers. Four utterances are taken from each speaker. Thus, a total of 200 test utterances taken outside of the training database is used in testing. Recognition experiments are carried out on the Japanese newspaper dictation task with a 20 K vocabulary. The language model is a standard word trigram model. The acoustic model is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total. A summary of the system specification is shown in Table I. We experimented using two reverberant conditions: $T_{60} = 200$ ms. and $T_{60} = 600$ ms. Reverberant training data were prepared by convolving the clean database with the measured ($\boldsymbol{T1}$) and estimated ($\boldsymbol{T2}$) RIRs. The latter is explained in the previous section. The test data are collected using a real recording ($\boldsymbol{E0}$) (1.5 m distance away from the microphone), filtering the clean speech with measured and estimated RIRs ($\boldsymbol{E1}$ and $\boldsymbol{E2}$), respectively. When collecting real recording data ($\boldsymbol{E0}$), we used movable panels built with absorption material in order to realize different reverberant effects within the room. The acoustic characteristics is altered by reconfiguring the movable panels. Thus, training and testing conditions are not the same.

For multiband spectral subtraction, we use a total number of bands $M = 5$, which is consistent to that of the former work [8]–[10]. We previously found [8], [9] that the improvement in recognition performance saturates at $M = 5$ and further increasing the number of bands only increased optimization time without significant recognition performance improvement. Moreover, we preliminarily investigated ordering of the spectral bands in the incremental optimization method. The difference

TABLE II
RECOGNITION RESULTS USING MEASURED RIR FOR TRAINING. (EVALUATION CONDITIONS: E0 REAL RECORDING; E1 CLEAN CONVOLVED WITH MEASURED RIR; E2 CLEAN CONVOLVED WITH ESTIMATED RIR)

	200msec			600msec		
Methods	***E*0**	***E*1**	***E*2**	***E*0**	***E*1**	***E*2**
(A) No processing (clean model)	65.7 %	67.3 %	60.1 %	21.4 %	22.8 %	15.3 %
(B) No processing (reverb model)	75.4 %	76.9 %	69.6%	32.1 %	33.8 %	28.7 %
(C) Multi-step LPC	80.9 %	81.4 %	77.1 %	56.8 %	58.4 %	54.0 %
(D) MMSE	80.1 %	81.3 %	76.4 %	54.2 %	55.6 %	50.3 %
(E) MMSE (decoding with proposed method)	**81.0** %	**81.9** %	**77.5** %	**57.1** %	**58.9** %	**54.7** %
(F) Batch (training only)	81.7 %	82.6 %	80.0 %	61.5 %	62.3 %	60.6 %
(G) Batch (training/decoding)	**82.8** %	**83.4** %	**81.6** %	**62.8** %	**63.5** %	**61.6** %
(H) Incremental (training only)	83.3 %	83.9 %	82.2 %	64.9 %	66.1 %	63.4 %
(I) Incremental (training/decoding)	**85.0** %	**85.7** %	**83.9** %	**66.2** %	**67.5** %	**65.3** %

TABLE III
RECOGNITION RESULTS USING ESTIMATED RIR FOR TRAINING. (EVALUATION CONDITIONS: E0 REAL RECORDING; E1 CLEAN CONVOLVED WITH MEASURED RIR; E2 CLEAN CONVOLVED WITH ESTIMATED RIR)

	200msec			600msec		
Methods	***E*0**	***E*1**	***E*2**	***E*0**	***E*1**	***E*2**
(A) No processing (clean model)	65.7 %	67.3 %	60.1 %	21.4 %	22.8 %	15.3 %
(B) No processing (reverb model)	72.2 %	73.5 %	78.8 %	30.3 %	31.6 %	39.5 %
(C) Multi-step LPC	78.0 %	79.3 %	81.4 %	54.5 %	58.2 %	59.7 %
(D) MMSE	**77.4** %	**78.1** %	**81.3** %	**51.8** %	**52.6** %	**56.9** %
(E) MMSE (decoding with proposed method)	**78.2** %	**79.4** %	**81.7** %	**54.9** %	**58.5** %	**60.1** %
(F) Batch (training only)	81.3 %	80.2 %	81.8 %	60.3 %	61.6 %	63.5 %
(G) Batch (training/decoding)	**82.4** %	**81.5** %	**83.3** %	**61.7** %	**62.6** %	**64.7** %
(H) Incremental (training only)	83.1 %	82.0 %	83.8 %	64.2 %	63.4 %	66.1 %
(I) Incremental (training/decoding)	**84.5** %	**82.9** %	**85.3** %	**65.7** %	**64.3** %	**66.8** %

in recognition performance after all bands are optimized in different orders is not statistically significant.

B. Basic Recognition Performance

The basic recognition performance of the proposed method is shown in Tables II and III, where the measured and the estimated RIRs are used to convolve the database in synthesizing the reverberant training data, respectively. In addition to the real reverberant data $\boldsymbol{E0}$, we also tested with a synthetically generated data derived by filtering the test utterances with the measured RIR $\boldsymbol{E1}$, and also generated data by filtering the test utterances with the estimated RIR $\boldsymbol{E2}$.

In Tables II and III, (A) is the performance when the reverberant test data is not processed at all (no dereverberation) using a clean acoustic model. (B) is the result when both testing and training are unprocessed. (C) is the performance when using multistep LPC [11]. (D) is the result when using the conventional MMSE-based approach while (E) is when combining the MMSE-based approach together with the GMM-based scale factor selection of the proposed method during decoding. (F) and (H) are the results of the proposed batch and incremental methods, respectively. It is confirmed that the proposed method is more effective than the conventional MMSE-based dereverberation technique.

In (E), (G), and (I), we can see the performance is further improved when optimization is also applied in the decoding process. Thus, optimizing dereverberation in both the acoustic modeling phase and decoding phase results in a synergetic effect in improving recognition accuracy. The performance of the proposed method is consistent for both real recording and synthetic reverberant test data for all of the three categories $\boldsymbol{E0}$–$\boldsymbol{E2}$.

We note that when evaluating the baselines in (A)–(E) the variation in recognition performance is larger among $\boldsymbol{E0}$–$\boldsymbol{E2}$. This effect suggests that mismatch is evident especially with the estimation of RIR. However, this becomes not an issue in (F)–(I) when using the proposed ASR-based optimization as manifested by the insignificant variation of recognition performance in $\boldsymbol{E0}$–$\boldsymbol{E2}$. The embedded retraining and the optimization of the scale factors during actual decoding minimize this effect.

Comparing Tables II and III, the degradation by using the estimated RIR is much smaller with the proposed methods for actual recording data $\boldsymbol{E0}$. Thus, the RIR estimation is effective for the proposed method.

C. Evaluation With Acoustic Model Adaptation

Model adaptation is used to address mismatch between training and testing conditions. Two popular adaptation schemes in ASR are the maximum *a priori* (MAP) [27] and the maximum-likelihood linear regression (MLLR) [28], [29]. In MAP [27], prior information on the training data is incorporated in the adaptation process. This is effective in dealing with sparse data. The MLLR adaptation approach [28], [29] estimates linear transformations or groups of model parameters to maximize the likelihood of the adaptation data.

We applied these supervised adaptation schemes (MAP and MLLR) to the proposed method. In this case, we execute an iterative MAP and MLLR, and in each iteration we optimize the dereverberation parameters using the 50 adaptation utterances. In the conventional iterative MAP and MLLR, only the model parameters are updated in every iteration, and the observation (adaptation) data is kept constant [30]. When expanding the proposed approach to MAP and MLLR, we also update the adaptation data using the proposed parameter optimization based on the acoustic model likelihood. In our experiment, we iterated five times, and in every iteration the adaptation data is updated.

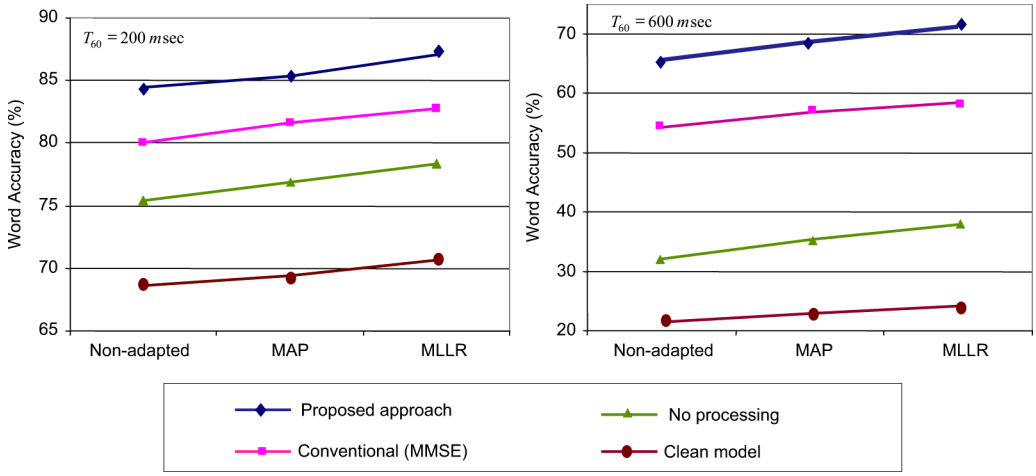

Fig. 5. Results on adaptation.

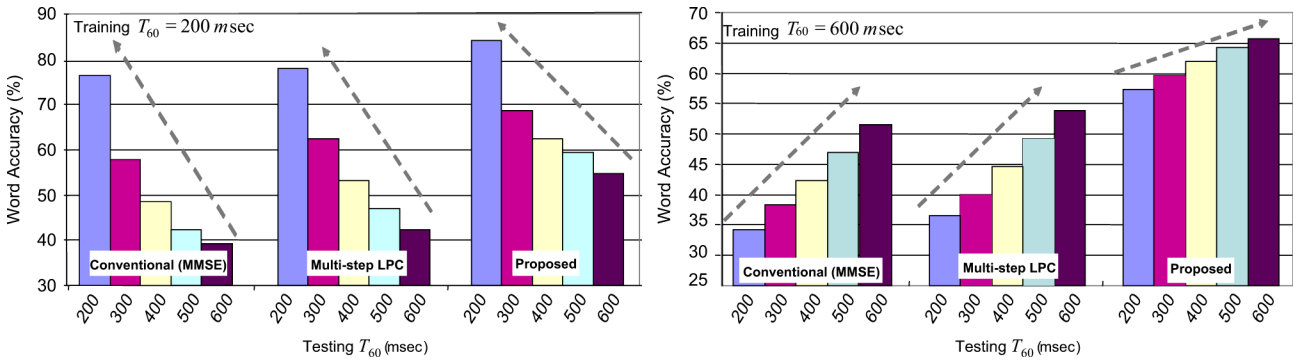

Fig. 6. Test for robustness against mismatch in reverberant conditions.

We note that in the conventional adaptation approach, the model does not have any bearing as to how the adaptation data is dereverberated for the next model update. Fig. 5 demonstrates that the proposed method (incremental (training/decoding) is effective in conjunction with adaptation, especially with MLLR, and the advantage over the conventional method is maintained after the adaptation.

D. Robustness Against Mismatch

We have shown that the proposed method works well with both synthetic and real data as shown in Tables II and III. The automatic RIR generator is shown to be effective in estimating T_{60} and used to replace the actual RIR measurement, which was a precondition of the conventional MMSE approach [8]–[10].

It is important to check whether the proposed method is robust to mismatch of the reverberant environment during speech recognition. For this evaluation, we prepared synthetic reverberant test data with different of T_{60} values, because it is difficult to obtain real reverberant data with a variety of T_{60}, and our previous results show consistency in real and synthetic reverberant data. In Fig. 6, two models are optimized for T_{60} of 200 ms and 600 ms, respectively. The proposed method performs better than the conventional approach as manifested by smaller changes in the drop of recognition performance across all of the test data with different T_{60}. When test data of different T_{60} are given to the system, the multistep LPC shows more robust performance than the conventional MMSE approach in mismatched conditions. However, the proposed method outperforms both the conventional MMSE and the multistep LPC as manifested by smaller changes in the drop of recognition performance across all of the test data with different T_{60}.

VII. Conclusion

We have presented an approach that performs dereverberation in relation with the likelihood of the back-end ASR system, covering both training and decoding phases. Dereverberation parameters are optimized based on the likelihood of HMM, and the dereverberation process is tightly integrated with acoustic model training. To further minimize the mismatch between training and actual testing conditions, we have implemented an additional optimization through fast selection of optimal dereverberation parameters used to process the actual reverberant data. The synergetic effect of the processes during training and decoding phases is confirmed, realizing an overall improvement of the recognition performance. The proposed method is effective in conjunction with the acoustic model adaptation. Moreover, we have shown that the proposed method is robust to mismatched reverberant environment conditions.

We have also shown that the use of RIR estimation works consistently well with the proposed approach in the adopted dereverberation scheme. In this paper, we described the scheme with the spectral subtraction-based method, but in theory, the proposed approach is applicable to other dereverberation techniques.

REFERENCES

[1] P. Naylor and N. Gaubitch, "Speech dereverberation," in *Proc. IWAENC*, 2005, pp. 173–176.
[2] Y. Huang, J. Benesty, and J. Chen, "Speech acquisition and enhancement in a reverberant, cocktail-party-like environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP*, 2008, vol. V, pp. 25–28.
[3] G. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Appl. Signal Process.*, vol. E80-A, pp. 804–808, 1997.
[4] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, p. 62, 2007.
[5] H. Attias, J. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press, 2001.
[6] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, Nov. 2008.
[7] B. Juang and L. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP'33, no. 6, pp. 1404–1413, Dec. 1985.
[8] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Distant-talking robust speech recognition using late reflection components of room impulse response," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 4581–4584.
[9] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Fast dereverberation for hands-free speech recognition," in *Proc. IEEE Workshop HSCMA*, 2008, pp. 140–143.
[10] R. Gomez, J. Even, H. Saruwatari, and K. Shikano, "Rapid unsupervised speaker adaptation robust in reverberant environment conditions," in *Proc. Interspeech*, 2008, pp. 1309–1312.
[11] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, vol. I, pp. 817–820.
[12] M. Seltzer, "Speech-recognizer-based filter optimization for microphone array processing," *IEEE Signal Process. Lett.*, vol. 10, no. 3, pp. 69–71, Mar. 2003.
[13] M. Seltzer and R. Stern, "Subband likelihood-maximizing beamforming for speech recognition in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2109–2121, Nov. 2006.
[14] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1996, pp. 353–356.
[15] D. Pye and P. C. Woodland, "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, pp. 1047–1050.
[16] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 10, no. 6, pp. 415–426, Aug. 2002.
[17] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Audio, Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
[18] W. Kim, S. Kang, and H. Ko, "Spectral subtraction based on phonetic dependency and masking effects," in *Proc. IEEE Visual Image Signal Process.*, Oct. 2000, vol. 147, pp. 423–427.
[19] P. Lockwood and J. Boudy, "Experiments with non-linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," *Speech Commun.*, vol. 11, no. 2–3, pp. 215–228, 1992.
[20] I. Soon, S. Koh, and C. Yeo, "Selective magnitude subtraction for speech enhancement," in *Proc. 4th Int. Conf./Exhib. High Perform. Comput. The Asia Pacific Region*, 2000, vol. 2, pp. 692–695.
[21] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Efficient blind dereverberation framework for automatic speech recognition," in *Proc. Interspeech*, 2005, pp. 3145–3148.
[22] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Amer.*, vol. 92, no. 2, pp. 1119–1123, Feb. 1995.
[23] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Commun.*, pp. 244–263, 2008.
[24] H. Kuttruff, *Room Acoustics*. London, U.K.: Spon Press, 2000.
[25] R. Gomez, T. Toda, H. Saruwatari, and K. Shikano, "Improving rapid unsupervised speaker adaptation based on HMM sufficient statistics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP*, 2008, vol. I, pp. 1001–1004.
[26] R. Gomez, T. Toda, H. Saruwatari, and K. Shikano, "Rapid unsupervised speaker adaptation using MLLR and speaker selection," in *Proc. Interspeech*, 2007, pp. 262–265.
[27] J. Gauvain and C. H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains," *Proc. IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
[28] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, no. 4, pp. 249–264, 1996.
[29] C. J. Leggeter and Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, pp. 171–185, 1995.
[30] HTK documentation. [Online]. Available: http://htk.eng.cam.ac.uk/docs/docs.shtml

Randy Gomez received the M.Eng.Sci. degree in electrical engineering from the University of New South Wales (UNSW), Sydney, Australia, in 2002 and the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology (Shikano Laboratory), Nara, Japan, in 2006.

His research interests include robust speech recognition, acoustic modeling, and adaptation. Currently, he is connected with the Academic Center for Computing and Media Studies (Kawahara Laboratory), Kyoto University, Japan.

Tatsuya Kawahara (M'91–SM'08) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1987, 1989, and 1995, respectively.

In 1990, he became a Research Associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Affiliated Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR, currently the National Institute of Information and Communications Technology (NICT). He has published more than 200 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been managing several speech-related projects in Japan including a free large-vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/).

Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. From 2003 to 2006, he was a member of the IEEE Signal Processing Society Speech Technical Committee. He was a general chair of the IEEE Automatic Speech Recognition and Understanding workshop (ASRU-2007).